BIOINFORMATICS

Vol. 18 no. 5 2002
Pages 689–696

Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions

Nela Zavaljevski[1], Fred J. Stevens[1] and Jaques Reifman[2,*]

[1]Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA and [2]US Army Medical Research and Materiel Command, 504 Scott Street, Fort Detrick, MD 21702, USA

Received on May 25, 2001; revised on November 31, 2001; accepted on December 6, 2001

ABSTRACT

Motivation: Data that characterize primary and tertiary structures of proteins are now accumulating at a rapid and accelerating rate and require automated computational tools to extract critical information relating amino acid changes with the spectrum of functionally attributes exhibited by a protein. We propose that immunoglobulin-type beta-domains, which are found in approximate 400 functionally distinct forms in humans alone, provide the immense genetic variation within limited conformational changes that might facilitate the development of new computational tools. As an initial step, we describe here an approach based on Support Vector Machine (SVM) technology to identify amino acid variations that contribute to the functional attribute of pathological self-assembly by some human antibody light chains produced during plasma cell diseases.

Results: We demonstrate that SVMs with selective kernel scaling are an effective tool in discriminating between benign and pathologic human immunoglobulin light chains. Initial results compare favorably against manual classification performed by experts and indicate the capability of SVMs to capture the underlying structure of the data. The data set consists of 70 proteins of human antibody $\kappa 1$ light chains, each represented by aligned sequences of 120 amino acids. We perform feature selection based on a first-order adaptive scaling algorithm, which confirms the importance of changes in certain amino acid positions and identifies other positions that are key in the characterization of protein function.

Contact: nelaz@ra.anl.gov; fstevens@anl.gov; jaques.reifman@amedd.army.mil

INTRODUCTION

Two recent publications (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001) documenting preliminary analysis of the human genome represent nothing less than a revolution in biological research, be it directed at fundamental issues, biomedical study, or biotech applications. The genomes of several hundred viruses, dozens of bacteria, and other organisms have been completed and many more projects, both in public and private venues, are aggressively underway.

Variations in DNA sequence and the predicted amino acid sequence derived from it can be correlated with discerned alterations in phenotype. However, an understanding of the basis for an observed correlation requires knowledge of the function or functions of the encoded protein, and insight into how single amino variations might directly or indirectly impair or enhance the functional and biophysical attributes of the gene product. This capability is dependent upon knowledge of the spatial relationships among all of the amino acids in the protein. This information requires knowledge of the three-dimensional structures of the protein itself, or a close structural homolog of the protein of interest. The importance of structural information to fully realize the implications and information content of the data generated by genomic sequencing has prompted the multinational initiation of structural genomics projects, which have been described in many recent publications, including Blundell and Mizuguchi (2000); Gershon (2000); Norvell and Machalek (2000), and Terwilliger (2000), to cite but a few. The goal of these projects is to complete the database of possible protein structures at a vastly accelerated rate.

The concurrent explosions of sequence and structural data have not been paralleled by an increase in the number of workers to correlate the data and extract meaningful new information and knowledge. This extraction process is critical to the way by which sequence and structure data contribute to both basic and applied research. Clearly, as neither the human genome project nor the emerging structural genomics program was possible without the introduction of extensive automation of experimental

*To whom correspondence should be addressed.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2001		**00-00-2001 to 00-00-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Medical Research and Materiel Command,504 Scott Street,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**8**	

methods, a similar challenge is presented by the need to merge sequence, structure, and functional data to construct an understanding of each protein system. Such capability can be accomplished through bioinformatics tools that relate extensive amino acid variation in a protein of known structure to achieve automated prediction of a functional attribute.

We recently completed an analysis of a database of human antibody light chain sequences from patients with plasma cell diseases and found that among the κ family of light chains we were able to correlate amyloid-forming capability with the presence of one or more of a limited set of structural 'risk factors' (Stevens, 1999, 2000). No such correlation has emerged among the λ family subset of data. This family, in general, appears to be intrinsically more amyloid-prone than the structurally homologous κ family. We, therefore, hypothesized that less additional destabilization through somatic mutation was required to increase the amyloidopotency of λ light chains (Stevens, 1999, 2000). This premise implies that new bioinformatic tools, such as might be found in approaches based on Support Vector Machines (SVMs), are needed to extract relevant information from the highly heterogeneous data. The first step in development of these tools is presented in this study, in which a simple SVM approach demonstrates accuracy of classification of κ data comparable to that previously achieved.

Support vector machines, a recently proposed supervised machine learning technique (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000), have been shown to be an effective bioinformatics tool in multiple areas of biological analysis. For example, SVMs have been used to functionally classify genes using knowledge-based analysis of microarray gene expression data (Brown *et al.*, 2000) and to classify microarray data in combination with gene selection (Furey *et al.*, 2000; Mukherjee *et al.*, 2000; Guyon *et al.*, 2001). They have been applied to infer gene functional classification from heterogeneous data sets consisting of DNA microarray expression measurements and phylogenetic profiles from whole-genome sequence comparisons (Pavlidis *et al.*, 2001). In addition, specially designed SVM kernels have also been used to recognize translation initiation sites in DNA (Zien *et al.*, 2000).

For protein analysis, SVMs have been less prevalent (Jaakkola *et al.*, 2000; Hua and Su, 2001; Ding and Dubchak, 2001). Instead, most applications to date have been based on Neural Network (NN) technology (Baldi and Brunak, 1998). For instance, Qian and Sejnowski (1988) and Holley and Karplus (1989) applied NNs for classification of protein secondary structure while Wu (1997) used them for prediction of protein tertiary structure.

In this study, we propose to use SVMs with selective kernel scaling for the classification of the κ family of human antibody light chains into benign or pathogenic categories and for the identification of markers, i.e. the selection of features, in the sequence of amino acids that are key discriminatory indicators. SVMs or other machine-learning tools have not yet been investigated for this problem. The selection of SVMs technology is driven primarily by their unique ability to construct predictive models with superior generalization power when the dimensionality of the data is high, i.e. the number of input features is large, and the number of observations available for developing (i.e. training) the model is limited. Their selection in this study is also attributed to their property of being capable of adapting to the problem at hand by including prior knowledge into the so-called kernel (mapping) function. We make use of this property to selectively scale the importance of amino acids in the sequence based on position variability at the germline level and position discriminatory power obtained through post-processing. In this work, we employ a version of the SVMlight code (Joachims, 1999) that we have modified to include selective kernel scaling. The original code is available at http://ais.gmd.de/~thorsten/svm_light.

SYSTEM AND METHODS

Data set

Historically, immunoglobulin light chains were one of the major early subjects of protein sequence determination. This occurrence was a consequence of at least two considerations. One, because of the cancer, multiple myeloma, large quantities of essentially pure, monoclonal protein could be obtained from patients as Bence Jones protein, so named for the physician who first described the material some 150 years ago. Two, these proteins provided the first glimpse of the structural origin of the diversity of antigen recognition by the immune system by demonstrating the diversity of the amino acid sequences found in the light chains, as well as heavy chains, of antibodies. At that time, it was not appreciated that in many patients the protein was a potential cause of death. To date, we have compiled a database of light-chain sequences from approximately 400 patients with plasma cell diseases. The data are essentially equally divided between the κ and λ classes of light chain. Because the earlier studies of Bence Jones proteins were oriented toward immunochemical interests, clinical data are available for only about half of the entries in the database. For the purposes of the work described here, we have limited our efforts to the same subset of data that was available for the previous study in which we identified four structural 'risk factors' that appear to reveal most amyloidogenic $\kappa 1$ light chains (Stevens, 1999, 2000).

The human light-chain database includes both κ and λ gene families encoded on separate chromosomes incor-

porating substantial amino acid variation. The κ family is represented by four major subgroups, of which the $\kappa 1$ subgroup is the most common. To further reduce inherited variation, primary sequences restricted to $\kappa 1$ light chains were extracted from the complete database and used for classification. Thus, our entire data set consists of 70 $\kappa 1$ light-chain proteins. Of those, six proteins were known to be benign, 33 were known to be pathogenic, i.e., from patients with amyloidosis, and 31 were of unknown pathology. Further analysis of the 31 unclassified proteins, including the use of the SVM classifier itself to identify misclassified proteins, allowed us to categorize 28 proteins into the benign class and the remaining three into the pathogenic class. Therefore, the final data set is almost equally divided (34/36) between the two classes, which avoids the construction of a class-biased classifier.

The SVM classifier receives a sequence of amino acids representing a protein as its input and predicts the class of the protein as its output. Because SVMs, as well as other machine-learning algorithms, use numerical values as inputs, they require the definition of encoding schemes. The encoding scheme for protein sequences can be rather involved and can greatly impact the performance of the classifier. One possibility is to encode each one of the 20 letters corresponding to the 20 amino acid types of a protein into a numerical scheme. In this case, each letter can be represented by a 20-dimensional binary vector indicating the presence of a particular amino acid or by a lower dimension vector based on the known physicochemical properties of each amino acid type (Baldi and Brunak, 1998). Here, we implement the latter scheme where each amino acid is represented by a set of seven physicochemical properties (Lohman *et al.*, 1994), hydrophobicity, hydrophilicity, polarity, volume, surface area, bulkiness, and refractivity, scaled to the $[-1, 1]$ interval. The primary structure of the κ light chains is aligned to 120 amino acid positions, which are, therefore, represented by 840 (120×7) input features to the SVM.

Support vector machines

Support vector machines are universal approximators that can be used to learn a variety of representations from training samples, and as such, are applicable to classification tasks and regression tasks (Vapnik, 1998). Their unique ability to develop models with superior generalization capabilities when the number of input features is large compared to the number of training samples is making this emerging technology the tool of choice among the various supervised learning algorithms, including NNs. Unlike NNs and other similar approaches where the number of model parameters that require estimation grow exponentially with the number of input features, the dimension of the SVM optimization problem is equal to the number of training samples. This unique capability affords their use for protein classification where the data sample is sparse and the dimension of the input features is large. The use of NNs, if attempted for this class of problems, would result in an overfitted model with very poor generalization capability.

When used for classification, SVMs map the input space into a higher-dimensional feature space that separates a given set of binary-labeled training data with an optimal hyperplane. The optimal hyperplane found by the SVM learning algorithm is the one that maximizes the separating margin between the binary classes of the training data and is defined by a relatively small number of M_S vectors in the input data set called support vectors. The motivation for mapping the data into a high-dimensional feature space is that linear decision boundaries constructed in the high-dimensional feature space correspond to nonlinear decision boundaries in the input space.

Given a training set of M samples or input vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_M\}$ with known class labels $\{y_1, y_2, \ldots, y_i, \ldots, y_M\}$, $y_i \in \{+1, -1\}$, a new data point $\boldsymbol{x}$ is assigned a label by the SVM according to the decision function

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{M_S} y_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \tag{1}$$

where

$$k(\boldsymbol{x}_i, \boldsymbol{x}) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}) \rangle \tag{2}$$

is the kernel function that defines the feature space, $\Phi(\boldsymbol{x})$ is a nonlinear mapping function from input space to feature space, $\langle \cdot, \cdot \rangle$ denotes an inner product, b is a bias value, and α_i are positive real numbers obtained by solving a Quadratic Programming (QP) problem that yield the maximal margin hyperplane (Vapnik, 1998).

A few important features of the algorithm should be noted. First, the number of free parameters of the QP problem is equal to the number of observations M in the training data. Second, the parameter α_i associated with each training point $\boldsymbol{x}_i$ expresses the strength with which that point is embedded in the decision function. It turns out, due to the nature of the QP problem, that only a subset M_S of the M points will be associated with non-zero α_i. These points are the support vectors, $\boldsymbol{x}_i$, $i = 1, 2, \ldots, M_S$, and are the points that lie closest to the separating hyperplane. Finally, the mapping function Φ need not be explicitly defined because the algorithm only requires the evaluation of the inner product in (2).

One of the most common kernels is the polynomial kernel

$$k_P(\boldsymbol{x}_i, \boldsymbol{x}_j) = (a + b\boldsymbol{x}_i \cdot \boldsymbol{x}_j)^d \tag{3}$$

where a, b, and d are real-valued constants. A special case of this kernel is the linear kernel obtained when $a = 0$ and $b = d = 1$. The kernel function, however,

can be customized to the problem at hand (Zien *et al.*, 2000). This unique feature of SVMs gives us the ability to implicitly incorporate prior knowledge, such as known physicochemical protein properties, into the mapping function by properly engineering the kernel function.

Selective kernel scaling

We customize the kernel function so that each component or input feature variables ℓ of the input vector i, x_i^ℓ, $\ell = 1, 2, \ldots, 840$ and $i = 1, 2, \ldots, M = 70$, could have a different scaling factor related to its importance to the classification problem. The modified kernel has the following form:

$$k_s(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\mathbf{S}\boldsymbol{x}_i, \mathbf{S}\boldsymbol{x}_j) \tag{4}$$

where $\mathbf{S}$ is a diagonal matrix of scaling factors. Here, we employ equal scaling for each group of seven properties representing each amino acid type and position-dependent scaling, based on the two schemes described below, to each of the 120 positions in the amino acid sequence.

1. Germline scaling

The first selective kernel scaling, termed germline scaling, is based on the assessment of the significance of the variability at specific positions in the amino acid sequence. All human κ light chains originate from a repertoire of about 14 inherited or germline genes. When these sequences are compared, 40 sites are invariant at the genetic level implying that evolution has conserved these positions for reasons that may include permissivity of fold, maintenance of stability, or contributions to one or more functional considerations. Other positions exhibit two or more alternative amino acids. We assume that positions that are conserved at the germline level would tend to have a higher probability of significantly affecting protein fold and/or stability, and therefore, lowered the weights assigned to positions of amino acid that exhibit variability at the germline level. Accordingly, the germline scaling factor for position n, $n = 1, 2, \ldots, 120$, is computed as $1/N(n)$, where $N(n)$ is the number of different amino acid types that appear at position n in the germline sequences.

2. Adaptive scaling

The second scheme, based on the post-processing of the classification problem, is adaptive and provides the means to iteratively modify the scaling factor of each input feature variable based on its affect or sensitivity on the classification. The Sensitivity Index (SI) of the classification function to a change in component ℓ of input feature vector i is to the first order of approximation given by (Evgeniou *et al.*, 2000)

$$\mathrm{SI}_\ell \approx \sum_{i=1}^{M_s} \left| \frac{\mathrm{d}f}{\mathrm{d}x_i^\ell} \right| = \sum_{i=1}^{M_s} \left| \sum_{j=1}^{M_s} \alpha_j y_j \frac{\mathrm{d}(k(\boldsymbol{x}_i, \boldsymbol{x}_j))}{\mathrm{d}x_i^\ell} \right|. \tag{5}$$

To compute the scaling factor of each group of seven input feature variables representing each amino acid in the sequence we add the SI in (5) over the seven properties, normalize the cumulative SI to the $[0, 1]$ interval and take the square root. When germline scaling is used in conjunction with adaptive scaling, the effective scaling factor for position n is taken as the square root of the product of $1/N(n)$ times the normalized cumulative SI value. The scaling factor provides a measure of the sensitivity of the classifier to perturbations of each amino acid position, and therefore, it is used here as a metric for feature selection—a needed capability in the identification of salient markers in the sequence of amino acids responsible for characterizing protein function.

Simulation results

Our method is tested in a number of simulation runs with the *SVMlight* code (Joachims, 1999) modified to include the scaling kernel schemes described above. The results are compared against a manual, heuristically performed classification approach (Stevens, 2000). Three measures of accuracy, classification error (E), recall (R), and precision (P), are used to assess the performance of the SVM classifier for the testing data

$$E = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \times 100\% \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{7}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{8}$$

where TP is the number of true positives, i.e., pathogenic proteins, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives. The classification error E provides an overall error measure, recall R provides a measure of the classification of the pathogenic proteins, and precision P provides a measure of the classification of the benign proteins.

The leave-one-out cross validation procedure is employed to estimate the classification accuracy of the 70-protein sequences. The leave-one-out procedure consists of removing from the training data one sample, constructing the decision rule on the basis of the remaining training data and then testing on the removed sample. In this fashion, one tests all $M = 70$ samples of the training data using 70 different decision rules (models) and computes the measures of accuracy in (6)–(8) based on the aggregate results over the 70 models.

The classification results using a linear kernel function with different position-dependent scaling schemes for the input features representing the sequence of amino acids are presented in Table 1. The entries without parenthesis

indicate results obtained by encoding the amino acids into the seven physicochemical properties discussed in the Section **System and methods** while the entries inside the parenthesis indicate results using six (polarity excluded) physicochemical properties. We first discuss the results obtained with the seven physicochemical properties. The direct application of the SVM algorithm with uniform scaling factors (i.e., all 840 (120×7) inputs equally weighted) achieved 72% classification accuracy (or alternatively, 28% classification error), 66% recall accuracy, and 75% precision accuracy. Adaptive scaling, in row two, employing the first-order SI in (5) improves all three measures of performance between 4–14%.

Adaptive scaling within the context of the leave-one-out cross validation procedure is performed as follows:

(1) Take 69 protein samples for training the SVM and leave one sample out for testing.

(2) Initialize the scaling factors for the ℓ inputs, $\ell = 1, 2, \ldots, 840$, based on uniform scaling or based on the germline sequence scaling.

(3) Train the SVM.

(4) Using (5), recompute the scaling factor for each of the ℓ input features and then aggregate and normalize the scaling factors for each group of seven inputs representing each amino acid.

(5) Repeat steps 3 and 4 until the improvement in the classification error for two consecutive iterations is $< \varepsilon$ (three to four iterations are usually sufficient).

(6) Test the remaining sample using the most recent SVM scaling factors.

(7) Repeat steps 1–6 for all 70 samples.

The classification results obtained by scaling the input features based on *a priori* knowledge about the significance of each specific amino acid mutation derived from conservation at the germline level is illustrated in the third row of Table 1. These results show very slight improvements over the uniform scaling case (row one). This may be consistent with the finding that most of these positions exhibit significant somatic variation and the fact that the relevant functional attribute, pathological aggregation, is generally a disease of advanced age and is not counter selected by evolution. The combination of the germline sequence scaling followed by adaptive scaling (row four) yields more significant improvements in each of the three measures of performance. The improvements range from 7–22% over the germline sequence scaling results presented in row three, confirming our hypothesis that the use of adaptive scaling results in an improved classifier with better generalization capabilities. In these simulations, the number of support vectors M_S for the various models ranged from 42 to 50 out the 70 available samples, indicating a substantial level of noise in the data.

Table 1. Leave-one-out classification accuracy based on several scaling schemes of the input features using seven and six (polarity excluded) physicochemical properties

Scaling scheme	Classification error (E) (%)	Recall (R) (%)	Precision (P) (%)
Uniform sequence scaling	28 (22)[a]	66 (72)	75 (81)
Uniform sequence scaling followed by adaptive scaling	24 (22)	72 (69)	78 (83)
Germline sequence scaling	27 (20)	72 (80)	74 (80)
Germline sequence scaling followed by adaptive scaling	21 (14)	77 (80)	80 (90)
Randomly assigned classes germline sequence scaling	55 (52)	64 (58)	44 (46)
Randomly assigned classes germline sequence scaling followed by adaptive scaling	54 (51)	35 (38)	42 (46)
Heuristic classification	15	94	79

[a]Entries in parenthesis indicate results obtained by encoding the amino acids into six physicochemical properties, excluding polarity.

To determine if the SVM was indeed learning the underlying structure of the data we repeated the simulations with randomly assigned labels for all 70 samples. The hypothesis being that, if the classifier was indeed learning the underlying structure of the data, as opposed to learning the structure of random data, its accuracy with randomly assigned labels should be about 50%. A large accuracy would indicate that the SVM is learning to explain noise. The results using germline sequence scaling are illustrated in the fifth row of Table 1, which indicate an overall classification error of 55%, clearly showing that the SVM is capable of capturing the underlying structure of the amino acid sequences. Reflecting an increase in the level of noise and a decrease in the information content of this data set, the range of support vectors increased to 57–61.

These simulations also serve to verify that the proposed adaptive scaling kernel algorithm does not result in an overfitted model that improves the explanation of the training data alone without improving the generalization of the classification model. When adaptive scaling was employed to the samples with randomly assigned labels the classification error remained essentially unchanged while recall and precision decreased, see row six in Table 1. An increase in the classification accuracy would have indicated that adaptive scaling is forcing the model to learn the structure of random data.

We also investigate the sensitivity of the SVM classifier to the removal of a few key amino acids and physicochemical properties used to encode the amino acids. We

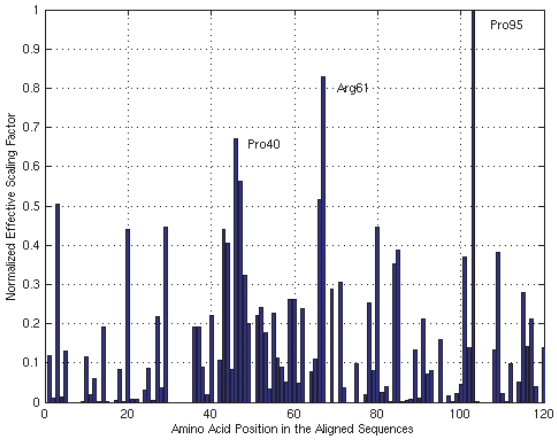

Fig. 1. Positional relevance of the amino acid sequence to classification accuracy based on mutations at the germline level and first-order SI.

removed each of the seven physicochemical properties, hydrophobicity, hydrophilicity, polarity, volume, surface area, bulkiness and refractivity, one at a time and repeated the simulations. The removal of bulkiness or polarity improved the classification results, indicating that these properties add noise to the model without providing any additional discriminatory information. The information provided by bulkiness can be considered to be redundant as both volume and surface area of the side chains can provide, approximately, the same information. The information provided by polarity, however, is generally important because electrostatic interactions are a long-range influence on both stability of the protein and its capacity to interact with other molecules. The entries inside the parenthesis in Table 1 illustrate the significant improvements obtained when polarity is excluded. Eighty-six percent classification accuracy is obtained when germline sequence scaling is combined with adaptive scaling (row four). This case is termed the 'best model' and is used as a base case in the comparisons below.

Figure 1 shows the normalized effective scaling factors aggregated over the entire data set for the 120 amino acid positions in the sequence. They can be employed as metrics for feature selection as they provide a measure of relevance of the contribution of each amino acid position to the classification. Hence, amino acid positions with large scaling factor values are key in discriminating between benign and pathogenic proteins.

Each of the three amino acid positions with large effective scaling factor values in Figure 1 has significant structural significance. For instance, position 61 is occupied by a highly conserved basic amino acid, Arg,

Table 2. Classification accuracy with key amino acids removed from the sequence

Positions removed	Classification error (E) (%)	Recall (R) (%)	Precision (P) (%)
40 (Pro40)	17	83	83
61 (Arg61)	31	72	68
95 (Pro95)	27	75	72
40, 61 and 95 (Pro40 + Arg61 + Pro95)	37	63	63
None	14	80	90

which forms a partially buried salt-bridge with the equally conserved acidic amino acid, Asp, located at position 82. This interaction contributes significant free energy to the stability of the protein. Any substitution of Arg61, including the nominally conservative substitution by the basic residue, Lys, is destabilizing and strongly associated with amyloid formation (Stevens *et al.*, 1995; Stevens, 2000). Pro95 is also highly conserved in human κ light chains. Disregarding the appearance of Ser at position 95 (encoded by an allele of the variable domain gene of κ1a), loss of Pro95 appears to be associated with light chain pathogenesis (Stevens, unpublished results). Mutations of position 40 were previously implicated (Stevens, 2000) as a significant contributor to amyloidosis. In particular, replacement of Pro with hydrophobic residues strongly correlated with fibril formation (Stevens *et al.*, 1995), probably as a consequence of two factors. First, although proline itself is significantly hydrophobic, replacement with leucine or isoleucine may increase the content of solvent accessible carbon. More importantly, however, is likely to be the loss of the backbone rigidity of proline, which when coupled with the glycine at position 41 provides the typical basis for tight beta-turn that occurs between beta strands C and C′.

The capability of the scaling factors to perform feature selection can be validated by retraining the SVM classifier with the three key amino acid positions removed. Table 2 compares the classification results when each of the three amino acid positions, Pro40, Arg61, and Pro95, is removed both independently and together with the previously obtained result of the 'best model,' which is repeated in the last row. The classification accuracy deteriorates when each of the three positions is independently removed, in particular, Pro95 and Arg61, and significantly deteriorates when all three positions are removed. These results clearly indicate that the scaling factors based on the SI in (5) provide a good mechanism for feature selection.

In all simulations, the SVM classification models based on a linear kernel were able to classify the 69 training data

points with 100% accuracy. This indicates that the data are linearly separable in the input space and the use of nonlinear kernels is not necessary and should be avoided to maintain model simplicity (Vapnik, 1998; Guyon *et al.*, 2001). Nevertheless, to investigate the stability of the proposed selective kernel scaling algorithms with regards to kernel forms we repeated the simulations using the polynomial kernel in (3). Numerous simulations involving various changes in the three parameters (a, b, and d) of the polynomial kernel resulted in identical classification accuracy as the ones obtained with the linear kernel in the first four rows in Table 1. The number of support vectors also remained almost unchanged ranging from 45 to 52 for the different models. Furthermore, comparisons of the effective scaling factors indicate only minimal variations around the distribution depicted in Figure 1. The importance of the same three amino acid positions, Pro40, Arg61, Pro95, was unmistakably distinguished in every simulation, which serves to demonstrate that the adaptive scaling algorithm is inherently stable, independent of kernel form.

The last row in Table 1 shows the results of a heuristic classification based on manual analysis of the data. The previous analysis (Stevens, 2000) was reduced to a subset of data to minimize confounding effects of inherited variation that may influence the significance of certain amino acid variations due to heterogenous structural contexts. Because all 70 samples were used to infer the heuristic classification rules, we cannot assess the generalization ability of the approach and perform a consistent comparison with the SVM results. Nonetheless, the heuristic approach provides for a semi-quantitative comparison, which indicates a favorable performance of the SVM algorithm. It should be noted that unlike the SVM algorithm, which is able to correctly separate the training data with no misclassifications, the heuristic approach misclassified 15% of the 'training' data.

CONCLUSIONS

Preliminary results demonstrate the ability of the SVM algorithm with selective kernel scaling to discriminate between benign and pathological immunoglobulins. The accuracy of the algorithm evaluated using the leave-one-out error estimate is promising and compares favorably with the accuracy obtained through manual heuristic classification performed by domain experts. Simulation tests where the protein labels, benign and pathological, are randomly assigned to the data are used to verify that the modified SVM is capable of capturing the underlying structure of the data. SVMs provide an effective inductive tool for developing protein classification models where the data is sparse and the dimensionality of the input features is large.

The use of selective kernel scaling increased the classification accuracy by as much as 36% without negatively impacting the generalization of the classification model. It also underscores the importance of incorporating *a priori* knowledge, such as the significance of conserved sites at the germline level, to selectively modify the kernel modeling function. Adaptive scaling based on first-order sensitivity analysis is shown to be a particular important tool for feature selection. It is able to confirm the importance of certain amino acid positions in the light chain, such as Arg61, and to identify new amino acid positions, such as Pro40 and Pro95, which contribute significantly to determining protein stability, previously shown to be the principal determinant of the pathological attribute of light chain pathology (Raffen *et al.*, 1999). The approach is stable in regards to kernel forms, providing the same performance improvements independent of the type of kernel used.

Future research will concentrate on two key topics. First, we will explore new algorithms for scaling the input feature variables and performing feature selection. Second, we will investigate ways in which the information content of the three-dimensional protein structure can be implicitly embedded in the scaling procedure. We believe that it is imperative to combine protein primary structure information with tertiary structure information to characterize the protein functional behavior—a critical feature ignored by simple analysis of strings of amino acid labels.

This point is illustrated by the observation that the inclusion of amino acid polarity had no significant effect in this study. The SVM method did successfully identify the importance of mutation Arg61 on protein pathology. This residue, which participates in a highly conserved ionic interaction with Asp82, has been observed frequently in pathogenic light chains. However, mutations that introduced the negatively charged Asp side chain at position 31, also highly correlated with pathogenicity (Stevens, 2000), did not contribute to the apparent significance of polarity. Clearly, the global contribution of many other changes in polar residues throughout the molecule may mask the importance of polarity changes at particular critical locations. This generalization should also apply to any other physicochemical feature when structural context is not incorporated.

ACKNOWLEDGEMENTS

The authors want to express their gratitude to T.Joachims for providing access to the SVMlight code and to the anonymous referees for useful comments and suggestions. The second author was supported by the US Department of Energy, Office of Biological and Environmental Research, under contract W-31-109-ENG-38 and by USPHS grants DK43757 and AG18001. The last author was supported in part by the Combat Casualty Care and

Military Operational Medicine Research Directorates of the US Army Medical Research and Materiel Command.

REFERENCES

Baldi,P. and Brunak,S. (1998) *Bioinformatics—the Machine Learning Approach*. MIT Press, Cambridge.

Blundell,T.L. and Mizuguchi,K. (2000) Structural genomics: an overview. *Prog. Biophys. Mol. Biol.*, **73**, 289–295.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Carrell,R.W. and Gooptu,B. (1998) Conformational changes and disease—serpins, prions, and Alzheimer's. *Curr. Opin. Struct. Biol.*, **8**, 799–809.

Cristianini,N. and Shawe-Taylor,J. (2000) *Support Vector Machines*. Cambridge, UK.

Ding,C.H.Q. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Evgeniou,T., Pontil,M., Papageorgiou,C. and Poggio,T. (2000) Image representation for object detection using kernel classifiers. *Fourth Asian Conference on Computer Vision*, January 9–11, Taipei, Taiwan, Paper ACCV-198, Poster Session.

Furey,T.S., Cristianini,N, Duffy,N., Bednarski,D.W., Schummer,M. and Huassler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Gershon,D. (2000) Structural genomics—from cottage industry to industrial revolution. *Nature*, **408**, 273–274.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Holley,L.H. and Karplus,M. (1989) Protein secondary structure prediction with a neural network. *Proc. Natl Acad. Sci. USA*, **86**, 152–165.

Hua,S. and Su,Z. (2001) A novel method of protein secondary structure prediction with segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **17**, 95–114.

Joachims,T. (1999) Making large scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge.

Lohman,R., Schneider,G., Nehrens,D. and Wrede,P. (1994) A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Sci.*, **3**, 1597–1601.

Mukherjee,S., Tamayo,P., Slonim,D., Verri,A., Golub,T., Mesirov,J.P. and Poggio,T. (2000) Support vector machine classification of microarray data, *MIT Report*, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, *AI Memo No 1667, CBCL Paper No. 182*.

Norvell,J.C. and Machalek,A.Z. (2000) Structural genomics programs at the US National Institute of General Medical Sciences. *Nature Struct. Biol.*, **7**, S931–S934.

Pavlidis,P., Weston,J., Cai,J. and Grundy,W.N. (2001) Gene functional classification from heterogeneous data. *Proceedings of the 5th International Conference on Computational Molecular Biology*, April 22–25, 2001, ACM Press, NY, Montreal, Canada, pp. 242–248.

Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Med. Biol.*, **202**, 865–881.

Raffen,R., Dieckman,L.J., Szpunar,M., Wunschl,C., Pokkuluri,P.R., Dave,P., Wilkins,S.P., Cai,X., Schiffer,M. and Stevens,F.J. (1999) Physicochemical consequences of amino acid variations that contribute to fibril formation by immunoglobulin light chains. *Protein Sci.*, **8**, 509–517.

Stevens,F.J., Solomon,A. and Schiffer,M. (1991) Bence Jones proteins: a powerful tool for fundamental study of protein chemistry and pathophysiology. *Biochemistry*, **30**, 6803–6805.

Stevens,F.J., Myatt,E.A., Chang,C.-H., Westholm,F.A., Eulitz,E., Weiss,D.T., Murphy,C., Solomon,A. and Schiffer,M. (1995) A molecular model for the formation of amyloid fibrils: immunoglobulin light chains. *Biochemistry*, **34**, 10 697–10 702.

Stevens,F.J. (1999) Protein structure, stability and conformational disease: human antibody light chains 1999. *Argonne National Laboratory Report, ANL/BIO/99-1*.

Stevens,F.J. (2000) Four structural risk factors identify most fibril-forming kappa light chains. *Amyloid: Int. J. Exp. Clin. Invest.*, **7**, 200–211.

Terwilliger,T.C. (2000) Structural genomics in North America. *Nature Struct. Biol.*, **7**, S935–S939.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wu,C.H. (1997) Artificial neural networks for molecular sequence analysis. *Comput. Chem.*, **21**, 237–256.

Zien,A., Ratsch,G., Mika,S., Scholkopf,B., Lengauer,T. and Muller,K.R. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 815–824.